DATA ANALYSIS PORTFOLIO

Abdul Azizur Rokhman

Hey, Let's get to know each other!

Hi, Introducing me Abdul Azizur Rokhman or can be called Aziz. I am an Informatics Engineering student from Universitas 17 Agustus 1945 Surabaya. Currently, I am interested in building a career in Data Science.

Through this portfolio, I would like to share my enthusiasm for Data Science.

This portfolio is a testament to my previous projects and data analysis skills. Hopefully with this portfolio, we can discuss possible opportunities in the future.


+62896-4483-9291


azizrokhman88@gmail.com


Abdul Azizur Rokhman


Aizuro

Educational Background

- Universitas 17 Agustus 1945 Surabaya (2022-Present)
  Informatics Engineering (GPA 3.61/4.00)

Skills

- Data Analysis
- English (Intermediate)

Advantage

- Leadership, Discipline, Time Management, Responsibility

MBTI (INFJ-A)

Organizational Experience

Himpunan Mahasiswa Teknik Informatika UNTAG Surabaya

(January 2024 - January 2025)

as Head of Entrepeneurship Division

- Sold shirts for the major and achieved the sales target of 80 shirts.
- Design and make shirts for major
- Design and make ID Card for committee

Project Background

This is the result of a project from the "Learning Data Analysis using Python" class organized by Dicoding Indonesia. The data in this project is Public E-Commerce Data which has a total of 9 Datasets, namely Customer Dataset, Geolocation Dataset, Order Items Dataset, Order Payments Dataset, Order Reviews Dataset, Orders Dataset, Product Category Name Dataset, Products Dataset, and Sellers Dataset with a total of 300,000+ data. To determine the steps that need to be taken in analyzing the data is to use SMART Questions.

S (Spesific)	M (Measurable)	A (Achievable)	R (Relevant)	T (Time-bound)
Which product categories have the lowest total order?	How much total revenue generated from the product category with the lowest total orders?	What promotional strategies can increase the revenue of underperforming products by 15% within the next quarter?	How does the price range of products impact their sales performance?	How can we increase the daily average number of orders over the next 3 months?

Data Wrangling

Data Gathering

- Customer Dataset

	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
0	06b8999e2fba1a1fbc88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0	14409	franca	SP
1	18955e83d337fd6b2def6b18a428ac77	290c77bc529b7ac935b93aa66c333dc3	9790	sao bernardo do campo	SP
2	4e7b3e00288586ebd08712fdd0374a03	060e732b5b29e8181a18229c7b0b2b5e	1151	sao paulo	SP
3	b2b6027bc5c5109e529d4dc6358b12c3	259dac757896d24d7702b9acbbff3f3c	8775	mogi das cruzes	SP
4	4f2d8ab171c80ec8364f7c12e35b23ad	345ecd01c38d18a9036ed96c73b8d066	13056	campinas	SP

- Geolocation Dataset

	geolocation_zip_code_prefix	geolocation_lat	geolocation_lng	geolocation_city	geolocation_state
0	1037	-23.545621	-46.639292	sao paulo	SP
1	1046	-23.546081	-46.644820	sao paulo	SP
2	1046	-23.546129	-46.642951	sao paulo	SP
3	1041	-23.544392	-46.639499	sao paulo	SP
4	1035	-23.541578	-46.641607	sao paulo	SP

- Order Items Dataset

	order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freight_value
0	00010242fe8c5a6d1ba2dd792cb16214	1	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	2017-09-19 09:45:35	58.90	13.29
1	00018f77f2f0320c557190d7a144bdd3	1	e5f2d52b802189ee658865ca93d83a8f	dd7ddc04e1b6c2c614352b383efe2d36	2017-05-03 11:05:13	239.90	19.93
2	000229ec398224ef6ca0657da4fc703e	1	c777355d18b72b67abbeef9df44fd0fd	5b51032eddd242adc84c38acab88f23d	2018-01-18 14:48:30	199.00	17.87
3	00024acbcdf0a6daa1e931b038114c75	1	7634da152a4610f1595efa32f14722fc	9d7a1d34a5052409006425275ba1c2b4	2018-08-15 10:10:18	12.99	12.79
4	00042b26cf59d7ce69dfabb4e55b4fd9	1	ac6c3623068f30de03045865e4e10089	df560393f3a51e74553ab94004ba5c87	2017-02-13 13:57:51	199.90	18.14

- Product Category Name Dataset

	product_category_name	product_category_name_english
0	beleza_saude	health_beauty
1	informatica_acessorios	computers_accessories
2	automotivo	auto
3	cama_mesa_banho	bed_bath_table
4	moveis_decoracao	furniture_decor

Order Payment Dataset

	order_id	payment_sequential	payment_type	payment_installments	payment_value
0	b81ef226f3fe1789b1e8b2acac839d17	1	credit_card	8	99.33
1	a9810da82917af2d9aefd1278f1dcfa0	1	credit_card	1	24.39
2	25e8ea4e93396b6fa0d3dd708e76c1bd	1	credit_card	1	65.71
3	ba78997921bbcdc1373bb41e913ab953	1	credit_card	8	107.78
4	42fdf880ba16b47b59251dd489d4441a	1	credit_card	2	128.45

Order Reviews Dataset

	review_id	order_id	review_score	review_comment_title	review_comment_message	review_creation_date	review_answer_timestamp
0	7bc2406110b926393aa56f80a40eba40	73fc7af87114b39712e6da79b0a377eb	4	NaN	NaN	2018-01-18 00:00:00	2018-01-18 21:46:59
1	80e641a11e56f04c1ad469d5645fdfde	a548910a1c6147796b98fdf73dbeba33	5	NaN	NaN	2018-03-10 00:00:00	2018-03-11 03:05:13
2	228ce5500dc1d8e020d8d1322874b6f0	f9e4b658b201a9f2ecdecbb34bed034b	5	NaN	NaN	2018-02-17 00:00:00	2018-02-18 14:36:24
3	e64fb393e7b32834bb789ff8bb30750e	658677c97b385a9be170737859d3511b	5	NaN	Recebi bem antes do prazo estipulado.	2017-04-21 00:00:00	2017-04-21 22:02:06
4	f7c4243c7fe1938f181bec41a392bdeb	8e6bfb81e283fa7e4f11123a3fb894f1	5	NaN	Parabéns lojas lannister adorei comprar pela I...	2018-03-01 00:00:00	2018-03-02 10:26:53

Orders Dataset

	order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date
0	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-10-04 19:55:00	2017-10-10 21:25:13	2017-10-18 00:00:00
1	53cdb2fc8bc7dce0b6741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	2018-07-24 20:41:37	2018-07-26 03:24:27	2018-07-26 14:31:00	2018-08-07 15:27:45	2018-08-13 00:00:00
2	47770eb9100c2d0c44946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	2018-08-08 08:38:49	2018-08-08 08:55:23	2018-08-08 13:50:00	2018-08-17 18:06:29	2018-09-04 00:00:00
3	949d5b44dbf5de918fe9c16f97b45f8a	f88197465ea7920adcdbec7375364d82	delivered	2017-11-18 19:28:06	2017-11-18 19:45:59	2017-11-22 13:39:59	2017-12-02 00:28:42	2017-12-15 00:00:00
4	ad21c59c0840e6cb83a9ceb5573f8159	8ab97904e6daea8866dbdbc4fb7aad2c	delivered	2018-02-13 21:18:39	2018-02-13 22:20:29	2018-02-14 19:46:34	2018-02-16 18:17:02	2018-02-26 00:00:00

Products Dataset

	product_id	product_category_name	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_length_cm	product_length_cm	product_height_cm	product_width_cm
0	1e9e8ef04dbcff4541ed26657ea517e5	perfumaria	40.0	287.0	1.0	225.0	16.0	16.0	10.0	14.0
1	3aa071139cb16b67ca9e5dea641aaa2f	artes	44.0	276.0	1.0	1000.0	30.0	30.0	18.0	20.0
2	96bd76ec8810374ed1b65e291975717f	esporte_lazer	46.0	250.0	1.0	154.0	18.0	18.0	9.0	15.0
3	cef67bcfe19066a932b7673e239eb23d	bebes	27.0	261.0	1.0	371.0	26.0	26.0	4.0	26.0
4	9dc1a7de274444849c219cff195d0b71	utilidades_domesticas	37.0	402.0	4.0	625.0	20.0	20.0	17.0	13.0

- Sellers Dataset

	seller_id	seller_zip_code_prefix	seller_city	seller_state
0	3442f8959a84dea7ee197c632cb2df15	13023	campinas	SP
1	d1b65fc7debc3361ea86b5f14c68d2e2	13844	mogi guacu	SP
2	ce3ad9de960102d0677a81f5d0bb7b2d	20031	rio de janeiro	RJ
3	c0f3eea2e14555b6faeea3dd58c1b1c3	4195	sao paulo	SP
4	51a04a8a6bdcb23deccc82b0b80742cf	12914	braganca paulista	SP

Insight

After seeing the columns of each dataset, it looks like I will later create 2 more datasets to make it easier for me to visualize the information I get. which one I will create a complete dataset for sales, and the other I will create a dataset for locations which I will use to find information about the locations of customers and sellers as well.

Data Wrangling

Data Cleaning

Problems that exist in each dataset and how to solve them:

- Customer Dataset

  there is no problem with any data

- Geolocation Dataset

  there is a lot of duplicated data, because the location points of each region are similar. I deleted all the duplicated data because later I only need one average location point for each city.

- Order Items Dataset

  there is a data type error in the shipping_limit_date column, which should be datetime but here the data type is object. I change the data type in the shipping_limit_date column from object to datetime

- Product Category Name Dataset

  there is no problem with any data

- Order Payment Dataset

  In the payment_value column, there is a minimum value of 0.00, it seems that this is caused by an unfinished payment. because the amount is only 0.2% of the total data, I just delete all data with a minimum value of 0.00.

- Order Reviews Dataset

  there are missing values on the review_comment_title and review_comment_message column, but I don't think that's a problem because it's just the comment not the review score given.

Data Wrangling

Data Cleaning

Problems that exist in each dataset and how to solve them:

- Orders Dataset

  There are data type errors in all date columns which should be datetime, and there are missing values in the order_approved_at, order_delivered_carrier_date, and order_delivered_customer_date columns. This is probably because the order was canceled by the buyer. The data type has been successfully changed to datetime, for the missing value, after I analyzed it, it seems that the average missing value data is canceled order data. The missing value problem in the orders dataset has been resolved by deleting all the missing value data because the average missing value data does not have important information for later analysis.

- Products Dataset

  there is a missing value in the product identity attribute, and in the product_weight_g attribute, there is a product that weighs 0g. I changed the missing value to “empty” for the product_category_name attribute, and 0 for product_name_lenght, product_description_lenght, and product_photos_qty.

- Sellers Dataset

  there is no problem with any data, it's just that in this dataset I added a customer column that contains the iteration number of each data to make it easier when visualizing.

Exploratory Data Analysis (EDA)

what I explore and what I discover:

- Customer Dataset

  In here, I explored the number of customers in each city/state

customer_city	customer_id
sao paulo	15540
rio de janeiro	6882
belo horizonte	2773
brasilia	2131
curitiba	1521
...	...
ibiara	1
rio espera	1
rio dos indios	1
rio dos cedros	1
lagoao	1

customer_state	customer_id
SP	41746
RJ	12852
MG	11635
RS	5466
PR	5045
SC	3637
BA	3380
DF	2140
ES	2033
GO	2020
PE	1652
CE	1336

  and the result is that the city/state that has the most customers is São Paulo/SP.

  Then, I also created a new column with the iteration number for each data to make it easier to identify when visualizing later.

	customer_id	customer	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
0	06b8999e2fba1a1fbc88172c00ba8bc7	1	861eff4711a542e4b93843c6dd7febb0	14409	franca	SP
1	18955e83d337fd6b2def6b18a428ac77	2	290c77bc529b7ac935b93aa66c333dc3	9790	sao bernardo do campo	SP
2	4e7b3e00288586ebd08712fdd0374a03	3	060e732b5b29e8181a18229c7b0b2b5e	1151	sao paulo	SP
3	b2b6027bc5c5109e529d4dc6358b12c3	4	259dac757896d24d7702b9acbbff3f3c	8775	mogi das cruzes	SP
4	4f2d8ab171c80ec8364f7c12e35b23ad	5	345ecd01c38d18a9036ed96c73b8d066	13056	campinas	SP

Exploratory Data Analysis (EDA)

what I explore and what I discover:

- Geolocation Dataset

  In here, I combine the geolocation dataset with the customer dataset and seller dataset, to create a visualization for intercity delivery speed.

	seller_city	seller_id	lat	lon	zip_code_prefix	seller
0	04482255	ceb7b4fb9401cd378de7886317ad1b47	-23.013945	-43.463020	22790	518
1	abadia de goias	c8143b3069f6746a77421b5ce30a450c	-16.767161	-49.438178	75345	2068
2	afonso claudio	2f4b9d112bfa44a214bc6cef085d17c8	-20.076743	-41.123686	29600	2154
3	aguas claras df	3a52d63a8f9daf5a28f3626d7eb9bd28	-15.832887	-48.029378	71900	2590
4	alambari	717b78b0950b51ed00b1471d858b0edc	-23.561571	-47.885101	18220	2116

	customer_city	customer_id	lat	lon	zip_code_prefix	customer
0	abadia dos dourados	9e01f714a2b3b8962c222cf2b74c20dc	-18.477752	-47.406319	38540	74530
1	abadiania	576d71ddb21b21763cfedce73b902180	-16.193718	-48.709452	72940	42308
2	abaete	08528824266cd0720658ff01df662b6a	-19.158364	-45.446897	35620	61141
3	abaetetuba	305f736d2711641a06f6a9c9bd837b00	-1.723644	-48.881349	68440	11201
4	abaiara	9723d86b12bdec025be4c43f71b0ba52	-7.355970	-39.043267	63240	69826

Exploratory Data Analysis (EDA)

what I explore and what I discover:

- Order Items Dataset

  In this dataset, I combining this dataset with the products dataset and the sellers dataset

	order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freight_value	product_category_name	seller
0	00010242fe8c5a6d1ba2dd792cb16214	1	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	2017-09-19 09:45:35	58.90	13.29	cool_stuff	514
1	00018f77f2f0320c557190d7a144bdd3	1	e5f2d52b802189ee658865ca93d83a8f	dd7ddc04e1b6c2c614352b383efe2d36	2017-05-03 11:05:13	239.90	19.93	pet_shop	472
2	000229ec398224ef6ca0657da4fc703e	1	c777355d18b72b67abbeef9df44fd0fd	5b51032eddd242adc84c38acab88f23d	2018-01-18 14:48:30	199.00	17.87	moveis_decoracao	1825
3	00024acbcdf0a6daa1e931b038114c75	1	7634da152a4610f1595efa32f14722fc	9d7a1d34a5052409006425275ba1c2b4	2018-08-15 10:10:18	12.99	12.79	perfumaria	2024
4	00042b26cf59d7ce69dfabb4e55b4fd9	1	ac6c3623068f30de03045865e4e10089	df560393f3a51e74553ab94004ba5c87	2017-02-13 13:57:51	199.90	18.14	ferramentas_jardim	1598

I explored the product performance and seller performance

product_category_name	product_id
cama_mesa_banho	3029
esporte_lazer	2867
moveis_decoracao	2657
beleza_saude	2444
utilidades_domesticas	2335

product_category_name	price
utilidades_domesticas	6735.00
pcs	6729.00
artes	6499.00
eletroportateis	4799.00
instrumentos_musicais	4399.87

seller	order_id
798	1854
2464	1806
1414	1706
475	1404
1874	1314

seller	price
2618	229472.63
902	222776.05
2464	200472.92
558	194042.03
1183	187923.89

the result is that the product that has the most sales is cama_mesa_banho, the highest price is utilidades_domesticas, and for the seller who has the most sales is the 798th seller, and the most revenue is the 2618th seller.

Exploratory Data Analysis (EDA)

what I explore and what I discover:

- Order Payment Dataset

  I explored the highest payment, what payment_type is widely used and the average customer makes purchases with how many payment_installments.

payment_type	order_id
credit_card	76505
boleto	19784
voucher	3866
debit_card	1528

payment_installments	order_id
1	49057
2	12389
3	10443
4	7088
10	5315
5	5234
8	4253
6	3916
7	1623

and the highest payment result is 13664.08, the type of payment that is widely used is credit_card, and most customers make payments only once.

Exploratory Data Analysis (EDA)

what I explore and what I discover:

- Order Reviews Dataset

  In here, I explore the product score

product_category_name	average_score	total_reviews
cds_dvds_musicais	4.642857	14
fashion_roupa_infanto_juvenil	4.500000	8
livros_interesse_geral	4.446266	549
construcao_ferramentas_ferramentas	4.444444	99
flores	4.419355	31
...	...	...
moveis_escritorio	3.493183	1687
pc_gamer	3.333333	9
portateis_cozinha_e_preparadores_de_alimentos	3.266667	15
fraldas_higiene	3.256410	39
seguros_e_servicos	2.500000	2

product_category_name	average_score	total_reviews
cama_mesa_banho	3.895663	11137
beleza_saude	4.142768	9645
esporte_lazer	4.107986	8640
moveis_decoracao	3.903493	8331
informatica_acessorios	3.930819	7849
...	...	...
cds_dvds_musicais	4.642857	14
la_cuisine	4.000000	13
pc_gamer	3.333333	9
fashion_roupa_infanto_juvenil	4.500000	8
seguros_e_servicos	2.500000	2

the result is, that the product that has the highest review score is musicais dvd cds but has a small total review of 14 only, for the product that has the most reviews is cama mesa banho but has a middle review score of 3.9, maybe from my vision the product that has the highest review score and the most total reviews is the beleza saude product with a review score of 4.1 with a total review of 9645.

Exploratory Data Analysis (EDA)

what I explore and what I discover:

- Orders Dataset

  In this dataset, I created a new column called delivery speed which contains the delivery speed (days), but I found something strange after I displayed the data, where the delivery speed data was minus.

	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date	delivery_speed
count	96461	96461	96461	96461	96461	96461.000000
mean	2018-01-01 23:53:26.642249216	2018-01-02 10:10:06.480142336	2018-01-05 05:21:04.508827392	2018-01-14 13:17:13.228102400	2018-01-25 17:33:14.236012544	9.292429
min	2016-09-15 12:16:38	2016-09-15 12:16:38	2016-10-08 10:34:01	2016-10-11 13:46:32	2016-10-04 00:00:00	-16.000000
25%	2017-09-14 09:28:28	2017-09-14 14:30:14	2017-09-18 16:52:19	2017-09-25 22:31:59	2017-10-05 00:00:00	4.000000
50%	2018-01-20 19:59:42	2018-01-22 13:49:00	2018-01-24 16:19:03	2018-02-02 19:50:56	2018-02-16 00:00:00	7.000000
75%	2018-05-05 18:33:24	2018-05-06 10:30:49	2018-05-08 14:33:00	2018-05-15 23:08:54	2018-05-28 00:00:00	12.000000
max	2018-08-29 15:00:37	2018-08-29 15:10:26	2018-09-11 19:48:28	2018-10-17 13:22:46	2018-10-25 00:00:00	205.000000
std	NaN	NaN	NaN	NaN	NaN	8.777252

  after I explored again, it turned out that there were 20 data that had a minus delivery speed, maybe it was due to human error or system error that was late in entering the delivery date, but I had overcome it by filling in the problematic delivery date with the date the order was approved, yes it might be different from the "original" delivery date but maybe not too far apart in days.

	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date	delivery_speed
count	96461	96461	96461	96461	96461	96461.000000
mean	2018-01-01 23:53:26.642249216	2018-01-02 10:10:06.480142336	2018-01-05 05:17:53.475674112	2018-01-14 13:17:13.228102400	2018-01-25 17:33:14.236012544	9.294658
min	2016-09-15 12:16:38	2016-09-15 12:16:38	2016-10-07 11:24:43	2016-10-11 13:46:32	2016-10-04 00:00:00	0.000000
25%	2017-09-14 09:28:28	2017-09-14 14:30:14	2017-09-18 16:52:19	2017-09-25 22:31:59	2017-10-05 00:00:00	4.000000
50%	2018-01-20 19:59:42	2018-01-22 13:49:00	2018-01-24 16:19:03	2018-02-02 19:50:56	2018-02-16 00:00:00	7.000000
75%	2018-05-05 18:33:24	2018-05-06 10:30:49	2018-05-08 14:33:00	2018-05-15 23:08:54	2018-05-28 00:00:00	12.000000
max	2018-08-29 15:00:37	2018-08-29 15:10:26	2018-09-11 19:48:28	2018-10-17 13:22:46	2018-10-25 00:00:00	205.000000
std	NaN	NaN	NaN	NaN	NaN	8.775443

After that, I merged this dataset with the seller dataset and the customer dataset.

	order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date
0	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-10-04 19:55:00	2017-10-10 21:25:13	2017-10-18
1	53cdb2fc8bc7dce0b6741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	2018-07-24 20:41:37	2018-07-26 03:24:27	2018-07-26 14:31:00	2018-08-07 15:27:45	2018-08-13
2	47770eb9100c2d0c44946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	2018-08-08 08:38:49	2018-08-08 08:55:23	2018-08-08 13:50:00	2018-08-17 18:06:29	2018-09-04
3	949d5b44dbf5de918fe9c16f97b45f8a	f88197465ea7920adcdbec7375364d82	delivered	2017-11-18 19:28:06	2017-11-18 19:45:59	2017-11-22 13:39:59	2017-12-02 00:28:42	2017-12-15
4	ad21c59c0840e6cb83a9ceb5573f8159	8ab97904e6daea8866dbdbc4fb7aad2c	delivered	2018-02-13 21:18:39	2018-02-13 22:20:29	2018-02-14 19:46:34	2018-02-16 18:17:02	2018-02-26

delivery_speed	customer	customer_city	customer_state	seller_id	seller	seller_city	seller_state
6.0	70297	sao paulo	SP	3504c0cb71d7fa48d967e0e4c94d59d9	560	maua	SP
12.0	77028	barreiras	BA	289cdb325fb7e7f891c38608bf9e0962	550	belo horizonte	SP
9.0	555	vianopolis	GO	4869f7a5dfa277a7dca6462dcf3b52b2	2618	guariba	SP
9.0	61082	sao goncalo do amarante	RN	66922902710d126a0e7d26b0e3805106	2990	belo horizonte	MG
2.0	67264	santo andre	SP	2c9e548be18521d1c43cde1c582c6de8	1497	mogi das cruzes	SP

Then I explored how fast inter-city delivery is and I also created a new column, the status column, which contains the customer's active status.

seller_city	customer_city	average_delivery_speed
guarulhos	marco	0.0
curitiba	parnamirim	0.0
borda da mata	rubim	0.0
guara	guaruja	0.0
videira	cotia	0.0
...	...	...
itajobi	perdizes	182.0
farroupilha	paulinia	186.0
aracatuba	aracaju	187.0
uberaba	lagarto	194.0
belo horizonte	montanha	195.0

status	customer_id
Active	96461
Non Active	2980

and the result, the fastest delivery is only a few hours not up to I day, and the slowest delivery is 195 days. Then for active customers there are 96461 and inactive ones are 2980, which can be concluded that active customers have more numbers than inactive customers.

Exploratory Data Analysis (EDA)

what I explore and what I discover:

- **All Dataset**
  This dataset is data that includes all information on sales data, by combining the Order Items dataset, Order dataset, Customer dataset, Seller dataset, and Order Reviews dataset.

	order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freight_value	product_category_name	seller
0	e481f51cbdc54678b7cc49136f2d6af7	1	87285b34884572647811a353c7ac498a	3504c0cb71d7fa48d967e0e4c94d59d9	2017-10-06 11:07:15	29.99	8.72	utilidades_domesticas	560
1	53cdb2fc8bc7dce0b6741e2150273451	1	595fac2a385ac33a80bd5114aec74eb8	289cdb325fb7e7f891c38608bf9e0962	2018-07-30 03:24:27	118.70	22.76	perfumaria	550
2	47770eb9100c2d0c44946d9cf07ec65d	1	aa4383b373c6aca5d8797843e5594415	4869f7a5dfa277a7dca6462dcf3b52b2	2018-08-13 08:55:23	159.90	19.22	automotivo	2618
3	949d5b44dbf5de918fe9c16f97b45f8a	1	d0b61bfb1de832b15ba9d266ca96e5b0	66922902710d126a0e7d26b0e3805106	2017-11-23 19:45:59	45.00	27.20	pet_shop	2990
4	ad21c59c0840e6cb83a9ceb5573f8159	1	65266b2da20d04dbe00c5c2d3bb7859e	2c9e548be18521d1c43cde1c582c6de8	2018-02-19 20:31:37	19.90	8.72	papelaria	1497

	customer_id	...	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date	delivery_speed	customer	customer_city	customer_state	seller_city	seller_state
9ef432eb6251297304e76186b10a928d		...	2017-10-02 11:07:15	2017-10-04 19:55:00	2017-10-10 21:25:13	2017-10-18	6.0	70297	sao paulo	SP	maua	SP
b0830fb4747a6c6d20dea0b8c802d7ef		...	2018-07-26 03:24:27	2018-07-26 14:31:00	2018-08-07 15:27:45	2018-08-13	12.0	77028	barreiras	BA	belo horizonte	SP
41ce2a54c0b03bf3443c3d931a367089		...	2018-08-08 08:55:23	2018-08-08 13:50:00	2018-08-17 18:06:29	2018-09-04	9.0	555	vianopolis	GO	guariba	SP
f88197465ea7920adcdbec7375364d82		...	2017-11-18 19:45:59	2017-11-22 13:39:59	2017-12-02 00:28:42	2017-12-15	9.0	61082	sao goncalo do amarante	RN	belo horizonte	MG
8ab97904e6daea8866dbdbc4fb7aad2c		...	2018-02-13 22:20:29	2018-02-14 19:46:34	2018-02-16 18:17:02	2018-02-26	2.0	67264	santo andre	SP	mogi das cruzes	SP

I explored what products have the highest revenue in each country.

customer_state	product_category_name	price
TO	utilidades_domesticas	926.17
	industria_comercio_e_negocios	69.00
	fashion_calcados	199.70
	fashion_bolsas_e_acessorios	426.59
	esporte_lazer	5481.24
...	...	...
AC	perfumaria	303.88
	relogios_presentes	1389.60
	telefonia	1103.96
	utilidades_domesticas	869.04
	artigos_de_natal	69.90

The result is that the country that has the highest total purchase price is in the country TO with the object that is bought at a high total price is esporte lazer.

I also explored what product categories have total order below 100 products.

	product_category_name_english	total_orders	price
61	security_and_services	2	183.29
29	fashion_childrens_clothes	7	89.99
11	cds_dvds_musicals	14	45.00
52	la_cuisine	14	140.00
3	arts_and_craftmanship	24	9.80
32	fashion_sport	29	49.90
46	home_comfort_2	30	12.90
35	flowers	33	49.90
23	diapers_and_hygiene	37	25.00
41	furniture_mattress_and_upholstery	37	110.00
55	music	38	220.00
58	party_supplies	42	257.60
27	fashio_female_clothing	45	54.90
9	books_imported	57	77.00
25	dvds_blu_ray	61	83.99
13	cine_photo	70	59.00
64	small_appliances_home_oven_and_coffee	73	1799.00
67	tablets_printing_image	83	69.90

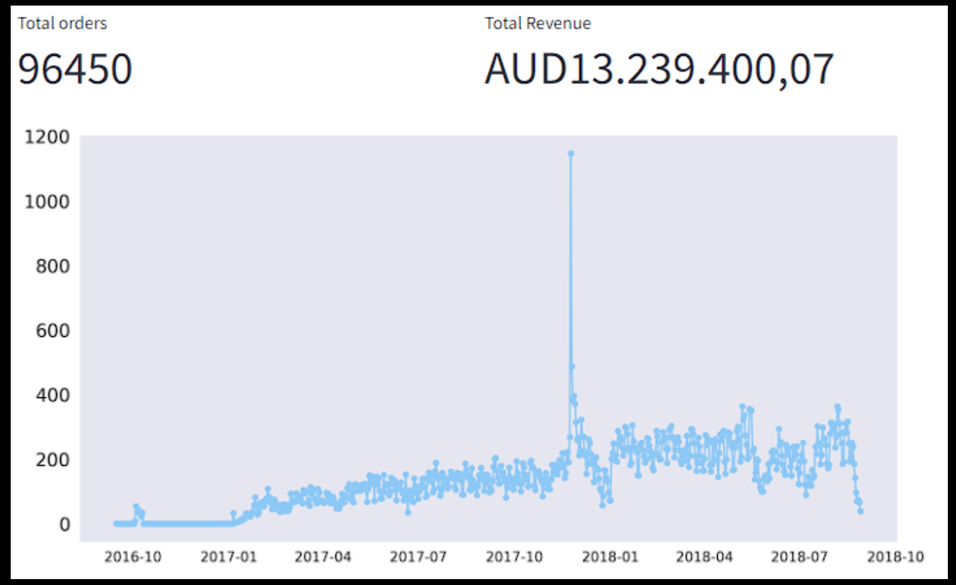
There are 18 product categories that have total orders under 100 products, and the product category that has the least orders is the security_and_services product category with a total order of only 2 products.

Data Visualization

After performing various steps in analyzing the data, this stage is to visualize the data in the dataset :

- Performance Overview

  In this visualization, contains how the performance of sales from the specified time



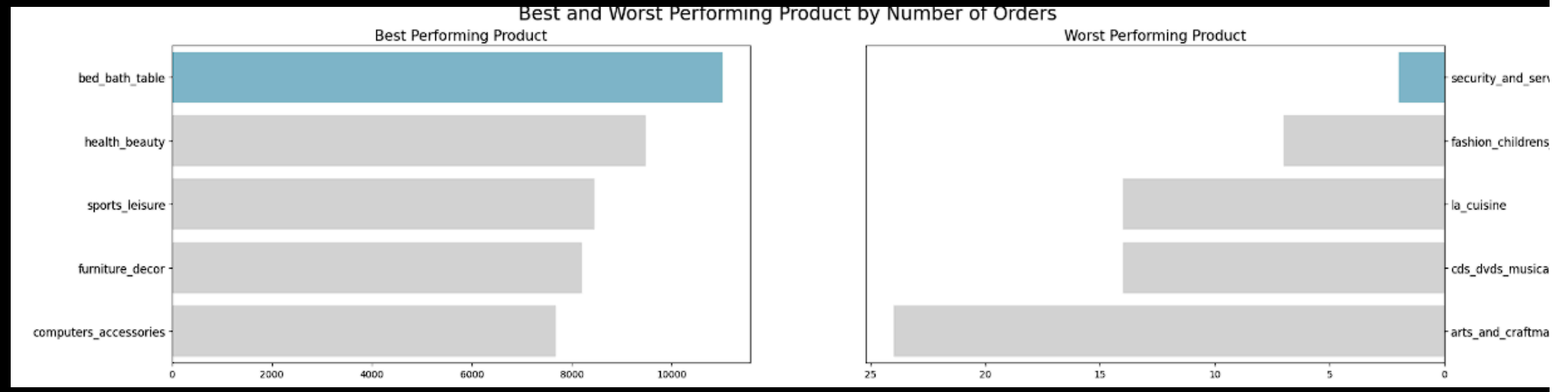
Over the entire time range, the total orders obtained reached 96450 with a total revenue of AUD 13.239.400,07. The distribution of total sales seems to be relatively increasing from month to month although it dropped slightly in the middle of 2018. It can be concluded that the overall sales graph is quite good, it just needs to be further identified why mid-2018 sales have declined slightly.

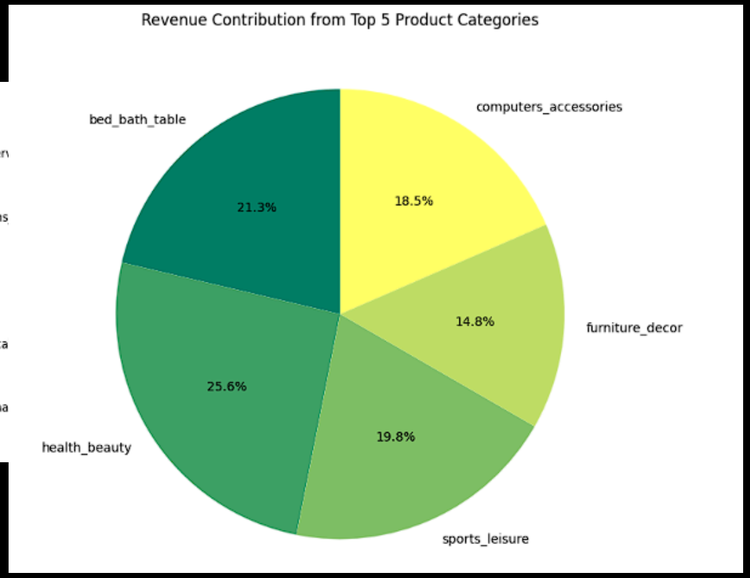
Data Visualization

After performing various steps in analyzing the data, this stage is to visualize the data in the dataset :

- Product Performance

  in this visualization, contains the performance of the product, such as its total sales and total revenue





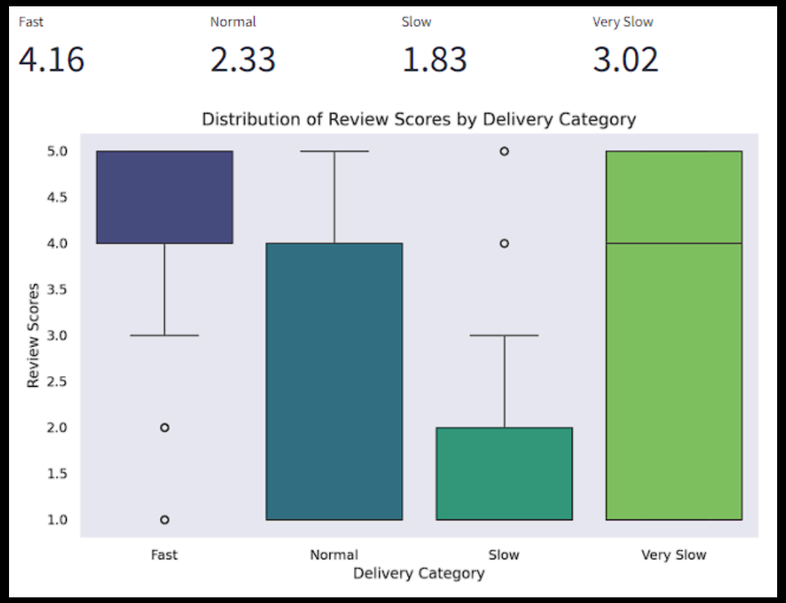
Based on the insight I got, the top 5 products that sold the most are bed_bath_table, health_beauty, sports_leisure, furniture_decor, and the 5th is computer_accessories, but here there is another insight that is quite interesting in the relationship between product sales and product contribution in revenue, which is that the product that sells the most does not necessarily have the most contribution to revenue, because the product that sells the most is bed_bath_table, while the product that has the most revenue contribution is health_beauty and bed_bath_table is in 2nd place, which shows that the best-selling product is not always the product that contributes to the most revenue, this can be caused by other factors such as price, and others.

Data Visualization

After performing various steps in analyzing the data, this stage is to visualize the data in the dataset :

- Distribution of Review Scores by Delivery Category

  in this visualization, contains the relationship between review score and delivery speed category



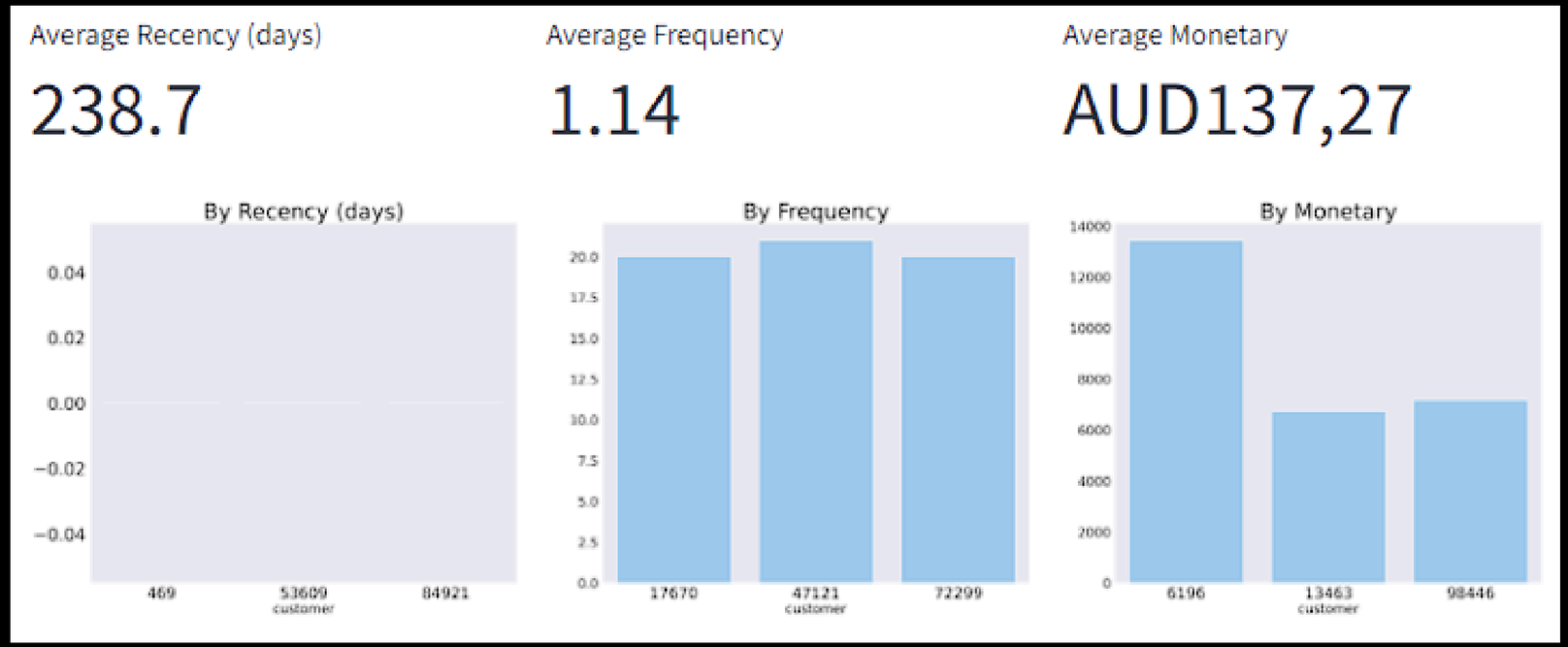
The fast category has a fairly good distribution of scores around 4.0 - 5.0, but there is some small data that is at a low score. For the normal category, the score distribution is fairly evenly distributed around 1.0 - 4.0 with the median around 4.0, which is certainly lower than the fast category. In the slow category, the score distribution is around 1.0 - 2.0 with the median around 2.0 but most of the data is in the lower score although there are some small data that are in the high score, even so the score in this slow category is lower than the normal category. Then for the very slow category, the distribution of scores is very evenly distributed from scores of 1.0 to 5.0 with the median being the same as the normal category, which is in the range of 4.0 but for the very slow category, there is quite a lot of data on high scores, which means that the scores in this slow category have higher scores than the normal category scores.

Data Visualization

After performing various steps in analyzing the data, this stage is to visualize the data in the dataset :

- **RFM Customer**

  in this visualization, contains recency, frequency, and monetary from customers from the last few times.



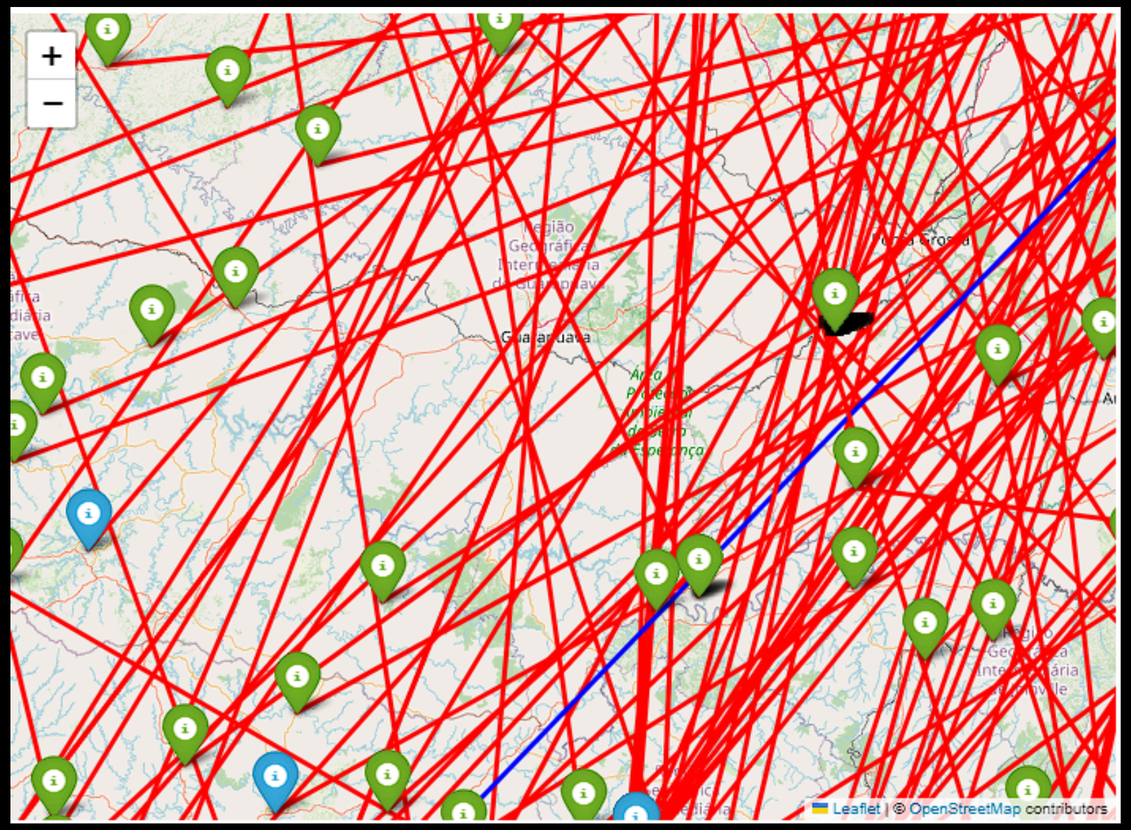
For the recency section, it turns out that many customers have been shopping within a few hours, this shows that there are still many customers who make purchases until now, although if you look at it, the average recency is at 238.7 days, maybe this happens because there are some customers who are no longer active. Then for the frequency, it turns out that here the frequency of customers shopping in the past is still quite a lot, with the top 3 in the range of 19 - 22 purchases. And finally for the monetary, there is also quite a lot of money earned with the most income in the 6196th customer with payments around 13,000 AUD. From this we can conclude that for customer activity from e-commerce there are still very many who are still actively buying

Data Visualization

After performing various steps in analyzing the data, this stage is to visualize the data in the dataset :

- Geospatial

  in this visualization, contains the delivery speed between cities



For this visualization, only for additions if you want to see the average delivery between cities, and if there are irregularities in the average delivery, you can immediately find out what is the cause of the long average delivery from this city to this city, if the cause is known, a follow-up of the problem will be carried out so that customers can be more happy with the quality of delivery that we provide.

Conclusion

Below is a review of the SMART Questions that have been created previously:

S (Spesific)	M (Measurable)	A (Achievable)	R (Relevant)	T (Time-bound)
Which product categories have the lowest total order?	How much total revenue generated from the product category with the lowest total orders?	What promotional strategies can increase the revenue of underperforming products by 15% within the next quarter?	How does the price range of products impact their sales performance?	How can we increase the daily average number of orders over the next 3 months?

after the results of the analysis that I did earlier, this is the conclusion:

There are a total of 18 product categories that have total orders below 100 (M), and the product category that has the lowest total orders is security_and_services with only 2 orders (S). If you look at the average price given to product categories that have low total orders, it is quite high, but it is possible that this is not the only factor that makes the total orders low, there are other factors such as the lack of need for the product category, then if I look at the product category, there are product categories that should be made into one, such as in the "diapers_and_hygienic" product category, it is quite broad in nature so maybe it can be made into one with other product categories that have similarities in the products sold (R). By conducting surveys for customer needs, then providing discount/bundle pricing of the less desirable products while promoting on social media and working with influencers to increase the visibility of the less desirable products, it is possible to increase the total orders of the less desirable products (A), as well as increase the daily total orders (T).

Tools

Python Language

Google Colab

Streamlit

Visual Studio Code

Links

Raw Data :
https://drive.google.com/file/d/1MsAjPM7oKtVfJL_wRpIqmCajtSGImdcK/view

Streamlit Cloud (Dashboard) :
https://app-sno8ztrquyngcdxqw9ojd6.streamlit.app/

Google Colab :
https://colab.research.google.com/drive/1WCZs6-DORSypWMRF5vstF8TWIYJnqAac?usp=sharing

Github :
https://github.com/Aizuro/Streamlit


Thank You!